Knowledge Extraction Using Dynamical Updating of Representation

ALDO DRAGONI
D.E.I.T., Università Politecnica delle Marche
Via Brecce Bianche
Ancona, Italy, 60131
dragon@inform.unian.it

GUIDO TASCINI
D.E.I.T., Università Politecnica delle Marche
Via Brecce Bianche
Ancona, Italy, 60131
tascini@inform.unian.it

LUIGI LELLA
D.E.I.T., Università Politecnica delle Marche
Via Brecce Bianche
Ancona, Italy, 60131
l.lella@inform.unian.it

WILLIAM GIORDANO
D.E.I.T., Università Politecnica delle Marche
Via Brecce Bianche
Ancona, Italy, 60131

Abstract

We present a system that extracts knowledge from the textual content of documents.

The acquired knowledge is represented through an associative network, that is dynamically updated by the integration of a contextualized structure representing the content of the new analysed document.

Grounded on the basis of "long term working memory" theory by W. Kintsch and K.A. Ericsson, our system makes use of a scale free graph model to update the final knowledge representation.

This knowledge acquisition system has been validated by first experimental results.

1 Introduction

From an historical perspective, four types of knowledge representation schemas are worth to be considered (W.Kintsch, 1998).

"Feature systems" (J.J. Katz, J.A. Fodor, 1963) have been developed in philosophy and linguistics and became very popular especially in psychology. This representation aimed at finding a limited set of basic semantic characteristics that, combined by means of particular composition rules, could express complex concepts. It was a very simple representation system but conceptual relations were not considered. Furthermore the defined features did not change with the context and the goals that had to be achieved.

"Associative networks" consider also semantic relations between concepts. Knowledge is represented by a network of concepts bounded by more or less strong associations. This formalism is bolstered by a lot of experimental data, for example by word priming experiments (D.E. Meyer, R.W. Schvaneveldt, 1971). But networks whose links are not labelled are not very expressive.

"Semantic networks" (A.M. Collins, M.R. Quillian, 1969) are an evolution of associative networks. Concepts continue to be symbolized by nodes, but these are linked by labeled arcs (IS-A, PART-OF etc.). In this way well ordered concept hierarchies can be defined and the hereditariness of properties is allowed.

"Schemas", "frames" and "scripts" are structures for coordinating concepts that belong to the same event or superstructure. Classical examples of these formalisms are the "room frame" of Minsky (M. Minsky, 1975) and the restaurant script of Schank and Abelson (R.C. Schank, R.P. Abelson, 1977).

The problem with these representation forms is that they are static. In fact human mind generates contextualized structures, that are adapted to the particular context of use.

"Networks of propositions" (or "knowledge nets", W.Kintsch, 1998) are an alternative formalism that combines and extends the advantages of the representation forms that have been introduced so far.

The predicate-argument schema can be considered as the fundamental linguistic unit especially in the representation of textual content. Atomic propositions consist of a relational term (the predicate) and one or more arguments.

Networks of propositions link these atomic propositions through weighted and not labeled arcs. According to this formalism the meaning of a node is given by its position in the net.

From a psychologic point of view only the nodes that are active (i.e. that are maintained in the working memory) contribute to specify the sense of a node. Hence the meaning of a concept is not permanent and fixed but is built every time in the working memory by the activation of a certain subset of propositions in the neighbour of the node that represents the concept. The context of use (objectives, accumulated experiences, emotional and situational state etc.) determines which nodes have to be activated.

For the definition of retrieval modalities Ericsson and Kintsch has introduced the concept of long term working memory (LTWM) (W.Kintsch, V.L. Patel, K.A.Ericsson, 1999). They noticed that some cognitive tasks, as textual comprehension, cannot be explained only using the concept of working memory. Given the strict limits of capacity of the short term memory (STM) and of the working memory (WM), tasks that require an enormous employment of resources cannot be carried out.

The theory of long term working memory specifies under which conditions the capacity of WM can be extended. The LTWM is involved only in the execution of well known tasks and actions, that belong to a particular cognitive domain that has been well experienced. In these cases the working memory can be subdivided in a short term part (STWM) that has a limited capacity and a LTWM that is a part of the long term memory represented by the network of propositions. The content of STWM automatically generates the LTWM. In particular objects present in the STWM are linked to other objects in the LTM by fixed and stable memory structures (retrieval cues).

2 Implementation of the Kintsch-Ericsson model

The approach of the network of propositions yielded two project problems. The creation of the LTWM and the activation of LTM nodes, i.e. the creation of the retrieval cues.

Kintsch has developed two methods for the definition of the LTWM.

The first, defined with Van Dijk (T.A. van Dijk, W. Kintsch, 1983), is a manual technique that starts from the propositions present in the text (micropropositions) and using some organizing rules arrives to the definition of macropropositions and macrostructures and even to the definition of LTWM.

The second is based on the latent semantic analysis (LSA) (T.K. Landauer, P.W. Foltz, D. Laham, 1998). This technique can infer, from the matrix of co-occurrence rates of the words, a semantic space that reflects the semantic relations between words and phrases. This space has typically 300-400 dimensions and allows to represent words, phrases and entire texts in a vectorial form. In this way the semantic relation between two vectors can be estimated by their cosine (a measure that according to Kintsch can be interpreted as a correlation coefficient).

This latter solution to the problem of the definition of LTWM puts a great and inevitable technical problem. How many objects must be retrieved from the semantic space for every word present in the text ? In some cases, when the textbase, i.e. the representation obtained directly from the text, is sufficiently expressed, the retrieval of knowledge from the LTM is not necessary. In other cases a correct comprehension of the text (or the relative situation model) requires the retrieval of knowledge from the LTM.

After the creation of the LTWM the integration process begins i.e. the activation of the nodes correspondent to the meaning of the phrase. Kintsch uses a diffusion of activation pocedure that is a simplified version of the one developed by McClelland and Rumelhart (J.L. McClelland, D.E. Rumelhart, 1986). Firstly an activation vector is defined whose elements are indexed over the nodes of LTWM. Any element's value is “1” or “0” depending on the presence or the absence of the corresponding node in the analyzed phrase (i.e. in the STWM). This vector is multiplied by the matrix of the correlation rates (the weights of the links of the LTWM) and the resulting vector is normalized. This becomes the new activation vector that must be multiplied again by the matrix of the correlation rates. This procedure goes on until the activation vector becomes stable. After the integration process, the irrelevant nodes are deactivated and only those that represent the situation model remain activated.

2.1 An alternative representation of the Kintsch-Ericsson model

The adoption of a network of propositions for the knowledge representation presents certainly great advantages in comparison with the classic formalisms. While semantic networks, frames and scripts organize knowledge in a more ordered and logical way, the networks of propositions are definitely more disorganized and chaotic, but present the not negligible advantage that are capable to vary dynamically not only in time, on the basis of the past experiences, but also on the basis of the perceived context.

But the technique worked out by Kintsch and Ericsson for the definition of LTWM presents some limits. Retrieving knowledge from the semantic space is only the first. Another problem is the evolution of the LTWM. The position occupied by a word in the LTWM is determined by the experience, i.e. its past use and this should be a lifetime experience. But this kind of knowledge cannot be reached practically and Kintsch resorts to the use of a dictionary for the definition of the semantic space that represents the LTWM.

Furthermore the construction-integration process does not always assure the semantic disambiguation of the analysed phrase (W.Kintsch, 1998).

The use of an external dictionary, as WordNet, (G. A. Miller, 1993) and of particular disambiguation procedures can overcome the last two limits.

Instead the first problem can be fully solved only by dropping the intermediate representation of the semantic space and by developing new methods for the direct formation of networks of concepts and propositions.

Let us describe now the system for the automatic acquisition of the knowledge that we developed on the basis of the LTWM model of Kintsch-Ericsson.

The lack of adequate textual parsers able to convert the paragraphs of a text in the correspondent atomic propositions has driven us to develop, at least in this initial phase of our project, simple dynamic models of associative networks.

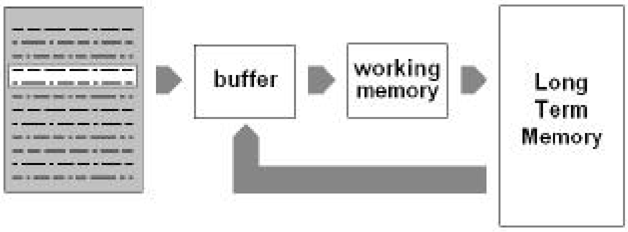

Figure 1: A possibile architecture of a system for the dynamical acquisition of knowledge from a repository of documents.

The part of the document that is analysed (the content of the buffer) must be codified on the basis of the context before being elaborated by the working memory block. The context represents the theme, the subject of the processed text and for its correct characterization not only the information present in the document must be considered, but also the one that can be retrieved from the structure representing the knowledge accumulated during the analysis of the previous documents presented to the system (Long Term Memory).

For the implementation of the working memory block, self organizing networks with suitable procedures for the labeling of their nodes could be used, but this solution requires a lot of computational time, especially for the analysis of entire repositories of documents.

So we considered alternative models based on the theory of scale free graphs (R.Albert, A.L.Barabasi, 2001) for the implementation of an associative network.

The graph theory dealed with regular graphs untill the 50s. Subsequently random graphs were introduced (P.Erdos, A.Renyi, 1959). They were the first simple forms of complex graphs that had ever been studied.

Their model started with a network made by N isolated nodes. Successively each pair of nodes could be connected with a probability p, leading to a graph having approximately pN(N-1)/2 links.

But this model was still far from real networks present in nature and artificial systems. So scientists defined other models characterized by an higher complexity level.

The actual models have three main features.

First their "small world" structure. That means there is a relatively short path between any two nodes (D.J.Watts, S.H.Strogatz, 1998).

Second their inherent tendency to cluster that is quantified by a coefficient that was introduced by Watts and Strogatz. Given a node i of k_i degree i.e. having ki edges which connect it to ki other nodes, if those make a cluster, they can establish $k_i(k_i-1)/2$ edges at best. The ratio between the actual number of edges and the maximum number gives the cluster coefficient of node i. The clustering coefficient of the whole network is the average of the all individual clustering coefficients. In a random graph the clustering coefficient is C = p. In real networks the clustering coefficient is much larger than p.

Actual graph models are also characterized by a particular degree distribution. While in a random graph the majority of the nodes have approximately the same degree close to the average degree, the degree distribution P(k) of a real network has a power-law tail $P(k) \sim k^{-?}$. For this reason these networks are called "scale free" (R.Albert, A.L.Barabasi, 2000).

Recently it has been found that human knowledge seems to be structured as a scale free graph (M.Steyvers, J.Tenenbaum, 2001). Representing words and concepts with nodes, some of these (hubs) establish much more links compared with the other ones.

In table 2 are reported the average shortest path length, the clustering coefficient and the power law exponent of two different types of semantic networks.

	Average path length	Clustering coefficient	Power law exponent
WordNet	10.56	0.0265	3.11
Roget Thesaurus	5.60	0.875	3.19

Table 1: General characteristics of some semantic networks.

This particular conformation seems to optimize the communication between nodes. Thanks to the presence of the hubs, every pair of nodes can be connected by a low number of links in comparison with a random network with the same dimensions. The definition and the eventual updating of a scale free network does not require a lot of time and the execution of particular processes, as the diffusion of the activation signal, is very fast.

The textual analysis is performed through the following steps.

The new text is analysed paragraph by paragraph. The buffer contains not only the words of the paragraph analysed, but also words retrieved from the long term memory using the diffusion of the activation procedure (the activation signal starts from the nodes in the LTM that represents the words in the paragraph). Theoretically, the buffer should contain also the words activated during the analysis of the previous paragraph, but this aspect has not been considered for its computational complexity. The buffer, the working memory and the activated part of the LTM block can be compared (but they are not the same structure) to the LTWM defined by Kintsch and Ericsson.

During the acquisition of the content of the paragraph a stoplist of words that must not be considered (as articles, pronouns etc.) is used.

For any word in the text, the paragraphs where it has appeared (or where it has been inserted after the retrieval procedure) are stored. When the entire text has been parsed and the data of all the N not filtered words have been memorized, the formation of the network of concepts in the working memory begins. The model adopted is similar to the one defined by Bianconi and Barabasi (G.Bianconi, A.Barabasi, 2001). The process starts with a net consisting of N disconnected nodes.

At every step t=1..N each node (associated to one of the N words) establishes a link with other M units (M=5). If j is the selected unit, the probability that this node establishes a link with the unit i is:

$$P_i = \frac{U_i k_i}{U_1 k_1 + \ldots + U_N k_N}$$

where k_i is the degree of the unit i [1], i.e. the number of links established by it, while U_i is the fitness value associated to the node, and it can be computed as the ratio between the number of paragraphs that contain both i and j and the number of paragraphs that contain either i or j.

LTM is an associative network that is updated with the content of the WM. Whenever a link of the WM corresponds to a link present in the LTM, the weight of this one is increased by “1”.

Example :

The WM links “Hemingway” to “writer”.

In the LTM “Hemingway” is linked to “writer” with weight “7” and to “story” with weight “4”.

In the updated LTM “Hemingway” is linked to “writer” with weight “8” and to “story” with weight “4” (unchanged).

To perform the diffusion of the activation signal all the weights must be normalized. In this case “Hemingway” must be linked to “writer” with weight 8/(8+4) and to “story” with weight 4/(8+4).

Since the scale free network that represents the content of the WM is used to update the content of LTM, this associative networks should take the form of a scale free graph. Unfortunately the modalities of evolution of the LTM does not allow the definition of a simple equivalent mathematic model, that is necessary to make useful previsions about its evolution.

In the scale free graph models proposed by literature at each temporal step M new nodes are added to the graph, with M defined beforehand. These M nodes generally establish M links with M old units of the network. In the system that we have developed, after the analysis of a new document the links related to an unknown number of nodes of the LTM network are updated on the basis of the content of the WM. This number depends on the analysed document because it is the number of the words that have not been filtered by the stoplist.

Another important difference with other scale free models presented in literature (S.N. Dorogovtsev, J.F.F. Mendes, 2001) is the particular fitness function that is used. This function does not depend on a single node but on the considered pair of nodes. If this value is choosen as proportional to the weights of the LTM associative network, the fitness value of a word is not constant but depends on the other word that could be linked to it. For example the noun “house” should present for the link with “door” a

[1] Each node is connected to itself by a loop.

fitness value greater than the ones presented for the links with “person” and “industry”.

3 Evaluation of the WM block

To test the validity of the scale free graph model adopted for the WM, we gave 100 files of the Reuters Corpus[2] as input to the system disabling the retrieval of information from the LTM.

Two versions of the model have been tested, one with bidirectional links and the other with directed links (in this case we considered $k_i = k_{i(IN)} + k_{i(OUT)}$).

In fig. 2 (http://www.deit.univpm.it/~dragoni /downloads/scale_free.jpg) an example of a network with bidirectional links is represented.

Please notice that the economic bias of the articles justifies the presence of hubs as “interest rate”, “economy”, etc., while other frequent words as “child”, “restaurant”, etc. establish less link with the others.

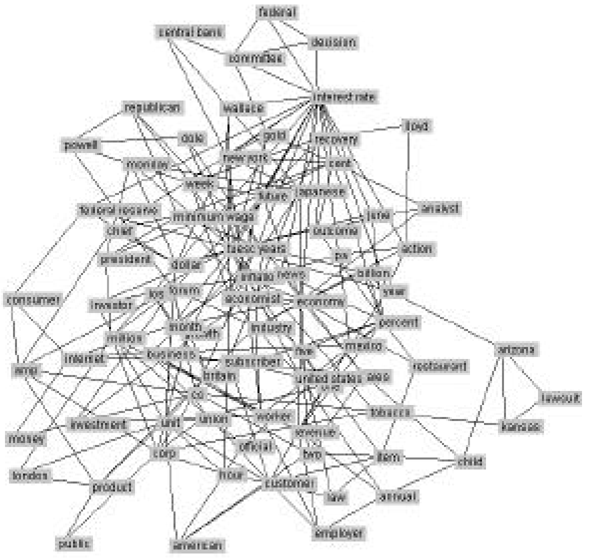

Figure 2: A network with bidirectional links obtained with the analysis of 100 files of the Reuters Corpus.

Fig.3 reports the average path length between each pair of nodes, the clustering coefficient and the degrees distribution of the nodes of the obtained networks.

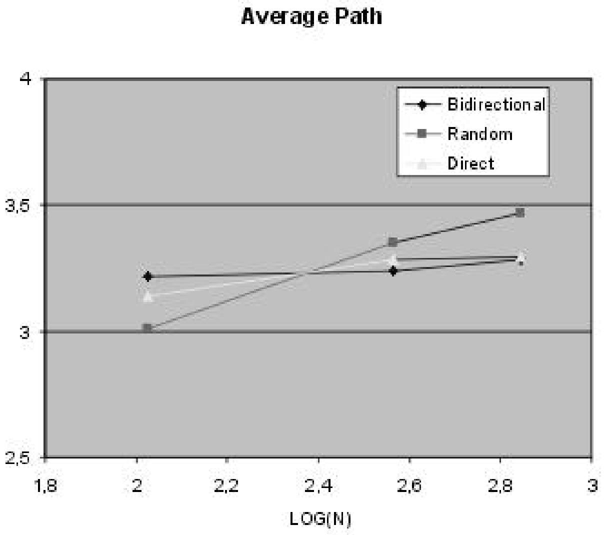

Figure 3: Comparison of average path lengths of different types of networks.

The tendency of the average path length is clear. The trend related to the random graphs, having the same dimensions of the considered scale free graphs, has an higher slope. This result confirms the one obtained by Bianconi and Barabasi reported in fig.4.

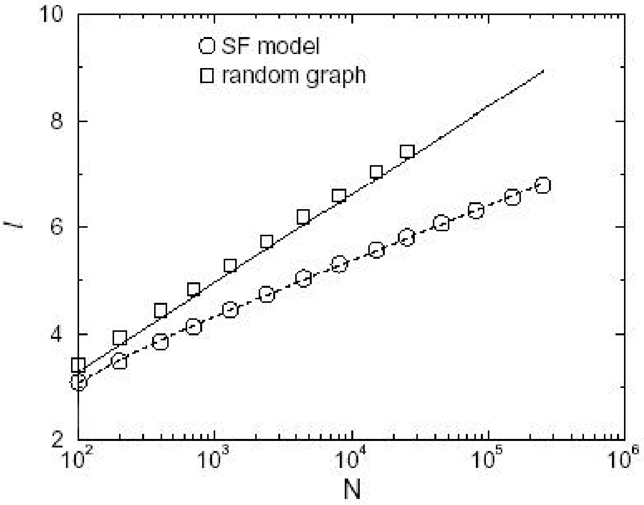

Figure 4: Comparison of average path lengths of different types of networks (Bianconi-Barabasi model).

Fig.5 shows that the clustering coefficient of the scale free graph model has an higher order of magnitude in comparison with the one computed for the random networks. Even this result is confirmed by the one obtained by Bianconi and Barabasi (fig.6).

[2] Reuters Corpus, Volume 1, English language, 1996-08-20 to 1997-08-19, http://about.reuters.com/ researchandstandards/corpus.

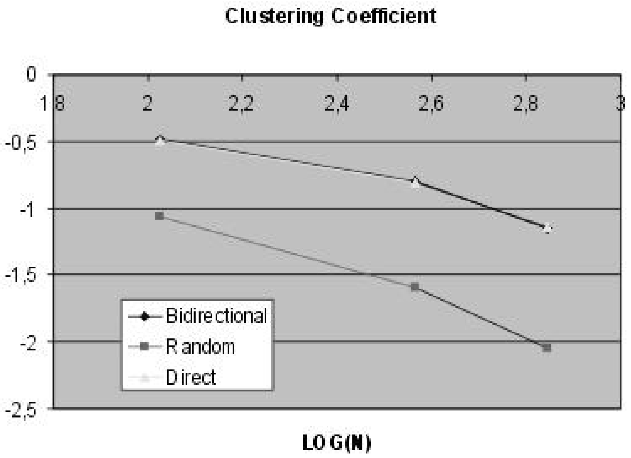

Figure 5: Comparison of clustering coefficients of different types of networks.

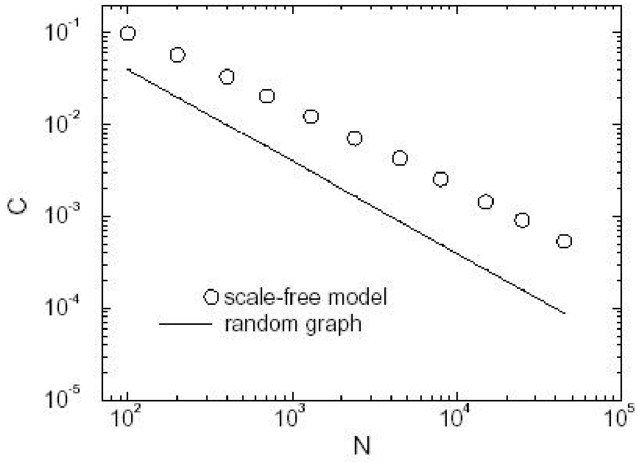

Figure 6: Comparison of clustering coefficients of different types of networks (Bianconi-Barabasi model).

Fig. 7 reports the degrees distribution of the graph with bidirectional links.

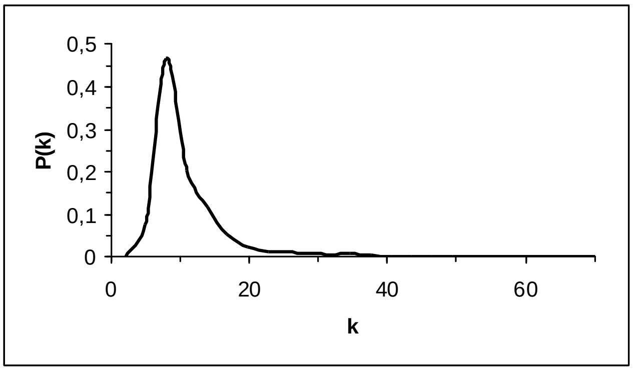

Figure 7: Degree distribution of a graph with M=5 and bidirectional links.

Fig. 8 highlights the trend by redrawing the graphic using the logarithmic coordinates.

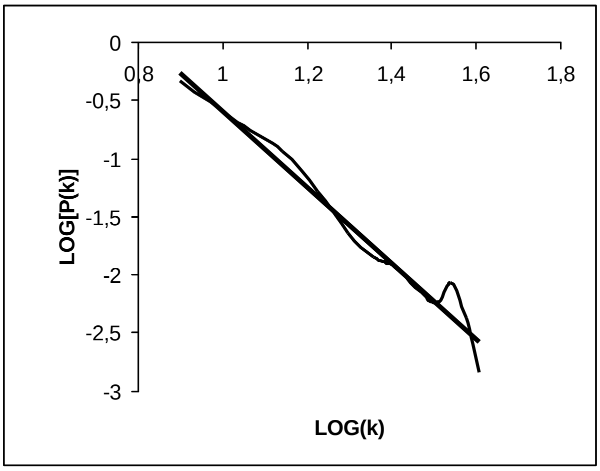

Figure 8: Previous graphic in logaritmic coordinates.

The degree distribution decays as P(k) ˜ k^{-G} with G = 3.2657.

The degree distribution of a graph with directed links is reported below.

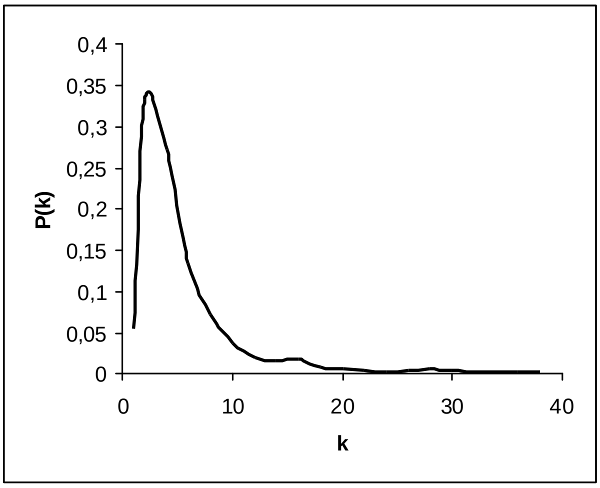

Figure 9: Degree distribution of a graph with M=5 and directed links.

Fig. 10 redraws the previous graphic using the logarithmic coordinates. The power law trend has a coefficient G = 2.3897.

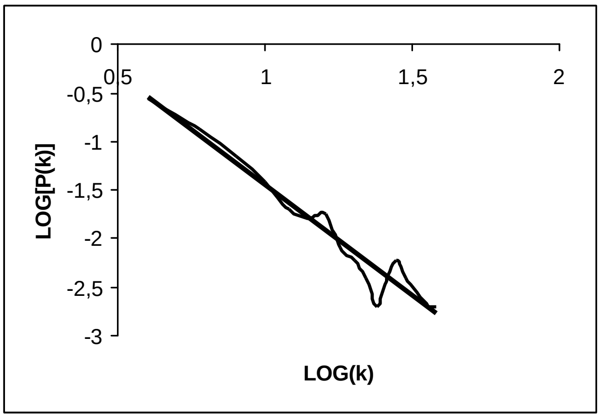

Figure 10: Degree distribution of a graph with M=5 and directed links.

4 Evaluation of the LTM block

In order to evaluate the learning capabilities of the system, we applied it on a medical article. The sections of the paper have been presented separately as independent texts regarding the same topic. This choice has been imposed by the necessity to enable also the retrieval of information from LTM.

As expected, the resulting LTM network was a typical scale-free graph (tab. 2).

M	Average path length	Average degree	Clustering coefficient
1	2.559	5.95	0.32290
2	2.499	6.50	0.33758
3	2.267	8.30	0.45428
4	2.255	9.50	0.43099
5	2.232	9.85	0.43151

Table 2: LTM with 40 nodes

The analysis has been repeated 30 times examining the coherence rate of each resulting LTM representation.

The coherence measure is based on a kind of transitivity assumption, i.e. if two concepts have similar relationships with other concepts, then the two concepts should be similar.

The coherence rate is obtained by correlating the LTM ratings given for each item in a pair with all of the other concepts[3]. Its value can be correctly computed only producing symmetric versions of the LTM data.

The average coherence rate was 0.45, indicating that the system has conceptualized the terms according to a precise inner schema.

To evaluate the correctness of this schema we are going to compare the obtained LTM representations with experimental data obtained from a group of human subjects. The subjects will be asked to read the same medical article examined by the system, assigning a rate of similarity to each pair of words that has been considered by the system. A Pathfinder analysis (R.W. Schvaneveldt, F.T. Durso, D.W. Dearholt, 1985.) will be performed on the relatedness matrices provided by human subjects and the LTM matrices in order to extract the so called “latent semantic”, i.e. other implicit relations between words. The obtained matrices will be compared using a similarity rate determined by the correspondence of links in the two types of networks.

5 Future work

Some important considerations can be made on the overall structure of the system.

The absence of an external feedback does not guarantee the correspondence between the LTM and the form of representation that must be modelled (the knowledge of an organization, the knowledge of a working group, the knowledge of a single user). A possible external feedback could be based on the evaluation of the performances of the system in the execution of particular tasks as the retrieval or the filtering of documents. For example the acceptance or the rejection of the documents selected by the system could be reflected in the updating modality of the LTM. In the first case the content of the WM could be used to strenghten the links in the LTM or to create new ones (as explained previously), in the second case the content of the WM could be used to weaken or delete the links in the LTM.

During the formation of the network in the WM the information about the weights of the links in LTM is not considered explicitly. Even if the weights can condition the retrieval of the information from the LTM, they could also modify the value of the fitness function used for the computation of the probability of the creation of new links in the WM.

Furthermore, the association of an age to the links of the LTM could guarantee more plasticity to its structure. Also the ages could be used in the computation of the fitness values, for example in accordance with the modalities suggested by Dorogovtsev (S.N. Dorogovtsev, J.F.F. Mendes, 2000).

We think that our knowledge acquisition system can be effectively used for the semantic disambiguation, that is the first phase of the analysis in the most recent systems for the

[3] All the operations described in this section are performed by the software PCKNOT 4.3, a product of Interlink Inc.

extraction of ontologies from texts (R. Navigli, P. Velardi, A. Gangemi, 2003).

As a further development, we are thinking of extracting from our representation form a simple taxonomy of concepts using techniques for the extraction of subsumption and equivalence relations. These techniques are based on the elaboration of the correlations between concepts expressed as fuzzy relations. A taxonomical representation can be considered as an important step towards the creation of an ontological representation. In this way our system could be used to model the user knowledge representing it in an ontological form.

6 Conclusions

A new system for the automatic acquisition of the knowledge has been presented. It is based on the concept of long term working memory developed by Kintsch and Ericsson.

The system updates an associative network (LTM) whose structure varies dynamically in time on the basis of the textual content of the analyzed documents. During the analysis of each new document the LTM can be queried by the simple procedure of the diffusion of the activation signal developed by Kintsch and Ericsson. In this way the context of the document can be easily and exactly identified.

To reduce the computational time we have implemented the WM block with a scale free graph model. The obtained network is used to update the content of the LTM.

Some analyses have been performed over the WM model developed. The results have confirmed that the network evolves as a scale free graph.

Also the LTM graphs seems to keep the scale free features, and their coherence rate indicates that the system conceptualizes the terms according to a precise inner schema.

Now we are considering alternative models for the WM that use much more information present in the LTM and that guarantee more plasticity to its structure. We are also going to compare the LTM graphs with the knowledge structures obtained by the Pathfinder analysis computed over the associations provided by a group of human subjects.

7 Acknowledgement

The authors are grateful to Prof. Ignazio Licata (Istituto di Cibernetica Non-Lineare per lo Studio dei Sistemi Complessi, Marsala(TP) - Italy) for helpful discussions, comments and criticisms.

References

R.Albert, A. Barabasi. 2000. *Topology of evolving networks: Local events and universality*. Phys. Rev. Lett. 85, p.5234.

R. Albert, A. Barabasi. 2001. *Statistical Mechanics of Complex Networks*. Rev. Mod. Phys., no.74, pp.47-97.

G. Bianconi, A. Barabasi. 2001. *Bose-Einstein Condensation in Complex Networks*. Phys. Rev. Lett., vol. 86, no. 24.

A.M. Collins, M.R. Quillian. 1969. *Retrieval from semantic memory*. Journal of Verbal Learning and Verbal Behaviour, 8, pp.240-247.

S.N. Dorogovtsev, J.F.F. Mendes. 2000. *Evolution of reference networks with aging*, arXiv: cond-mat/0001419.

S.N. Dorogovtsev, J.F.F. Mendes. 2001. *Evolution of networks.* arXiv: cond-mat/0106144, submitted to Adv. Phys.

P.Erdos, Renyi A.. 1959. *On Random Graphs*. Publ. Math. Debrecen 6, p. 290.

J.J. Katz, J.A. Fodor. 1963. *The structure of semantic theory*. Language, 39, pp.170-210.

W. Kintsch. 1998. *Comprehension. A Paradigm for Cognition*. Cambridge University Press.

W. Kintsch. 1998. *The Representation of Knowledge in Minds and Machines*. International Journal of Psychology, 33(6), pp.411-420.

W. Kintsch, V.L. Patel, K.A.Ericsson. 1999. *The role of long-term working memory in text comprehension*. Psychologia, 42, pp.186-198.

T.K. Landauer, P.W. Foltz, D. Laham. 1998. *An Introduction to Latent Semantic Analysis.* Discourse Processes, 25, pp.259-284.

J.L. McClelland, D.E. Rumelhart. 1986. *Parallel distributed processing*. Cambridge, MA: MIT Press.

D.E.Meyer, R.W. Schvaneveldt. 1971. *Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations*. Journal of Experimental Psychology, 90, pp.227-234.

G. A. Miller. 1993. *Five papers on WordNet*. Cognitive Science Laboratory Report 43.

M. Minsky. 1975. *A framework for representing knowledge*. In P.H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill.

R. Navigli, P. Velardi, A. Gangemi. 2003. *Ontology Learning and Its Application to Automated Terminology Translation*. IEEE Intelligent Systems, January/February 2003, pp. 22-31.

R.C. Schank, R.P. Abelson. 1977. *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Erlbaum.

R.W. Schvaneveldt, F.T. Durso, D.W. Dearholt. 1985. *Pathfinder: Scaling with network structures.* Memorandum in Computer and Cognitive Science, MCCS-85-9, Computing Research Laboratory. Las Cruces: New Mexico State University.

M. Steyvers, J. Tenenbaum. 2001. *The Large-Scale structure of Semantic Networks*. Working draft submitted to Cognitive Science.

T.A. van Dijk, W. Kintsch. 1983. *Strategies of discourse comprehension*. New York: Academic Press.

D.J. Watts, S.H. Strogatz. 1998. *Collective dynamics of 'small-world' networks*. Nature, vol. 393, pp. 440-442.